全国第四届研究生数学建模竞赛

题号 A

题 目 食品卫生安全保障体系数学模型及改进模型

摘 要：

本文通过建立人群食物摄入量模型、污染物分布模型、风险评估模型并建立了一套比较完整的膳食暴露评估数学模型。在食物摄入量模型中，采用分区设点进行抽样，食品分类进行统计的抽样调查方案，同时考虑人群性别、年龄、经济收入、劳动强度等因素。然后利用辅助信息增设虚拟样本来扩充样本数，利用多元线性回归技术得到食物摄入量模型的函数。采用聚类分析的方法解决局部样本概率分布与总体不一致的问题。在污染物分布模型中，对于 2%的小样本数据，我们采用 Bootstrap 方法多次再抽样以达到获得大样本的目的。并创新性的尝试引入了信道处理中的 Nakagami-m 分布逼近。风险评估模型中，以两种思路进行展开：一、直接利用上述两个模型中的抽样数据进行计算获得抽样样本中每个人的各种污染物日摄入量，然后对于超标部分的数据进行 PI 估计和 TE 估计。其中 TE 估计主要是利用极限值理论对超标数据进行尾部分布逼近，然后用 Pareto 法则对这一分布进行修正，得到与现实情况相似的概率分布函数，进而求出 99.999%右分位点。二、通过前两个模型的函数和概率分布，进行点评估、单一分布、双随机分布三种估计方法来得到相应的风险评估指标。点估计较简单，易于操作，能提供一定的参考信息，但精确度不高；单一分布精度较高，而且也能克服数据的不匹配的缺憾，但对于数据的随机性与不确定的应变能力不够强。双随机分布是一类较理想的估计法，当前两个模型的精度比较高时，这种估计能够很好地提供风险评估的各种指标。最后，由于数据来源有限，本文只进行了部分数值实验。

关键词：多元线性回归，聚类分析，Bootstrap，Nakagami-m 分布，Pareto 法则

参赛密码 _____

（由组委会填写）

参赛队号 1000401 参赛学校 北京交通大学

参赛队员姓名 余家新 王高阳 罗自炎

第一部分 模型概述

1.1 研究现状

近年来，我国重大食物中毒事件时有发生，给国家和个人都造成不小的损失。由卫生部公布的 2007 中国卫生统计提要数据可知，2006 年全国发生食物中毒共 1978 起，中毒人数共 31860 人，死亡人数 209 人[1]。为了评价全国食品安全状况和预警可能出现的食品安全事件，我们需要针对不同地区、不同人群、不同季节的各类食物摄入量和食品污染物含量等数据进行抽样，然后应用概率统计学和多元分析等知识建立食品卫生安全保障体系数学模型，即食物安全风险的膳食暴露数学评估模型。风险评估模型的输入变量是我们最新一时段所进行的有关膳食调查或食品检测数据，输出变量是该时段下，分别针对全国、某地区、某类食物的人均摄入污染物导致中毒的风险(概率)，同时显示对该时段下的食物风险评估(99.999%右分位点数值)。

根据现阶段的实际观察和欧美国家的做法,我们将根据我国实际国情建立膳食暴露数学评估模型，膳食暴露评估模型分为三个步骤来完成,即建立人群事物摄入量模型、建立污染物分布模型、建立风险评估模型(暴露评估概率模型)，如下图 1 所示。

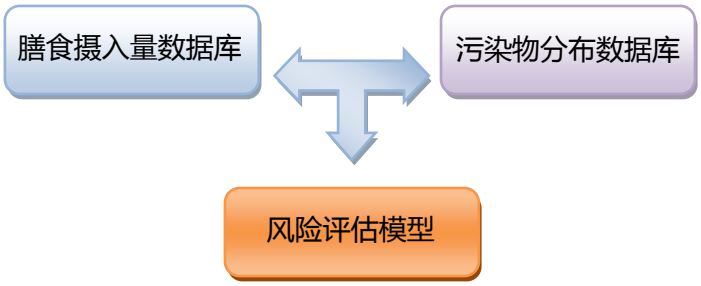

图 1 膳食暴露数学评估模型结构

第二部分 子模型 1—人群食物摄入量模型

2.1 模型基本假设

2.1.1 全国受调查人群的分类

将全国按照地域性划分为有限个区。每省为一个具体实施单位，下设多个调查点，如农村点和城市点。食物摄入量调查分别以住户、个体以及团体（如学校，工厂食堂等）为单位进行。以省、市、自治区计，每个点调查多个单位。按照各个地区的饮食习惯和消费水平的不同，可将中国分为南北各两个区共四个区，并在这四个区中有意识的选取能代表所在区域特征的省份来进行抽样调查。如下图 1 所示四个抽样的区域，在各个抽样的区域中，根据人口的分布、经济水平等因素按一定的比例进行抽样。

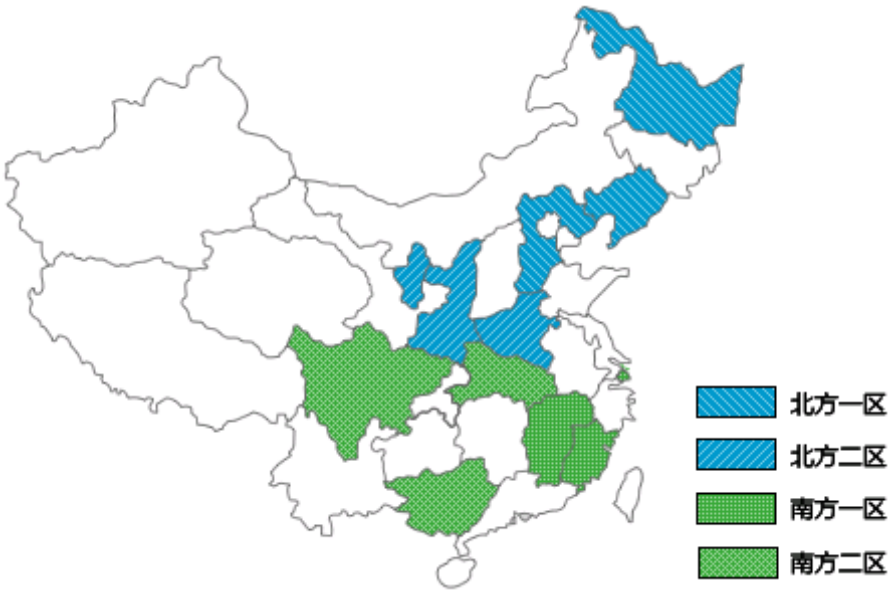

图 1 数据抽样区域划分

2.1.2 全国所检测食物的分类

食物的分类可以根据污染物模型中各种有化学污染物的分布特征来确定需要调查的食物种类和食物的分类。我们调查的食物优先考虑那些人们日常生活中摄入量比较大的，含某种或某些有害物质浓度或者受污染概率比较大的食物种类，而对于具有相同或者相似的危害的食物我们可以归为一类来进行统计。例如食品 1 和食品 2 中的危害成分相同且都具有较高的浓度，则可将这两种食品归为一类来统计。

结合我国地域特点，食物一般分类为米及米制品、面及面制品、其他谷类、豆类及豆制品、蔬菜、水果、猪肉、牛肉、家禽、其他肉类、奶及奶制品、蛋类、鱼虾类、坚果、蛋糕及淀粉类、食用油、盐、酒类和其他。

2.2 模型数据库的建立

2.2.1 数据的采集--抽样具体流程

人群食物摄入量调查数据表如表 1 所示。

表 1 食物摄入量调查数据表

姓名: 性别: 年龄: 体重: 劳动强度: 经济收入: 调查时间:								
日期	摄入食物(g)							
	大米	面食	其他谷类	豆类	素菜	水果	猪肉	…
总摄入量(g)								
平均每日摄入量(g)								

食物摄入量调查过程流程图如 2 所示。

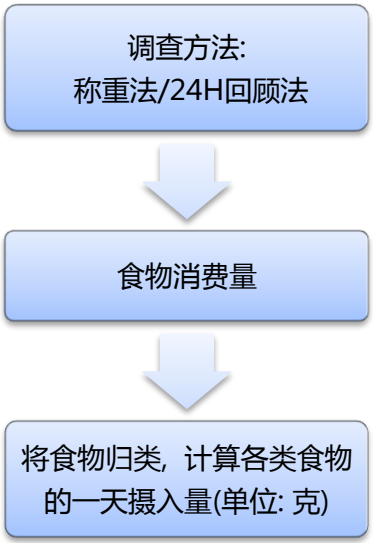

图 2 人群食物摄入量调查过程流程图

2.2.2 数据的采集--抽样方法与分类

常用于评估集体用餐单位、散居居民和个体膳食的调查方法有称重法、查账法、24 小时回顾法(询问法)、频率法和化学分析法。这五种方法各有其优点及局限性[2]，总结如表 2 所示。一般认为称重法能获得准确、可靠的食物摄入量，常把称重法作为“标准”以评价其他方法的准确性。针对不同的调查人群，采取不同的调查方式，如家庭或者个人一般采用称量法得到准确的食物摄入量，对于学校等大型的具有单一特征人群的团体可以采用查账法。

表 2 常用五种调查方法的应用范围和优缺点总结

	优点	缺点	应用
称重法	准确	费时费力，不适用大规模	家庭、个人、团体
查账法	简单易行,省时、人、物	时间短，不够准确，代表性有影响	账目清楚的机关部队学校
24 小时回顾法	简单易行,省时、人、物	主观，不太准确，回忆偏倚	家庭、个人
化学分析法	准确	费时，费力，费财	科研、治疗膳食
频率法	应答者负担轻，应答率高，经济、方便；可调查长期	量化不准确(偏高)，遗漏	个人，膳食习惯与某些慢性疾病的关系

我国在 20 世纪 80 和 90 年代多采用称重法进行膳食调查。1992 年第三次全国营养调查对 24h 回顾法进行改进，利用 24 小时回顾法得出的食物摄入量结合称重法的调味品摄人量取代传统称重法，并应用于 2002 年中国膳食状况调查。我们可以采取称重法和 24 小时回顾法结合的方法来进行膳食调查，使调查的全部工作量控制在可以承受的范围内。

2.3 模型的建立

2.3.1 多元线性回归建模

按照上述抽样方法对数据采集后，得到的食物摄入量测量值为Y，食物按照前面的分类分为n类，则此处变量Y为n维向量。$x_1,x_2,\cdots,x_m$分别为影响因子，包括年龄、性别、经济收入、季节和区域等等。

假设变量 Y 与 m 个变量$x_1,x_2,\cdots,x_m$的关系是线性关系。

$$Y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_mx_m+\varepsilon \tag{2.1}$$

其中，ε为随机项，且ε_i满足正态分布$N\left(0,\sigma^2\right)$。则记

$$Y=\begin{bmatrix}y_1\\y_2\\\vdots\\y_n\end{bmatrix},X=\begin{bmatrix}1&x_{11}&\cdots&x_{1m}\\1&x_{21}&\cdots&x_{2m}\\\vdots&\vdots&&\vdots\\1&x_{n1}&\cdots&x_{mm}\end{bmatrix}_{n\times(m+1)},\varepsilon=\begin{bmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_n\end{bmatrix},\beta=\begin{bmatrix}\beta_1\\\beta_2\\\vdots\\\beta_m\end{bmatrix} \tag{2.2}$$

则变量 Y 与变量 X 的关系可以化为

$$Y=X\beta+\varepsilon \tag{2.3}$$

求回归系数$\beta_0,\beta_1,\cdots,\beta_m$的最小二乘估计。作残差平方和

$$Q=\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1x_{i1}-\cdots-\beta_mx_{im}\right)^2 \tag{2.4}$$

求 Q 分别关于$\beta_0,\beta_1,\cdots,\beta_m$的一阶偏导数，并令它们等于零，得

$$\beta=\begin{bmatrix}\beta_1\\\beta_2\\\vdots\\\beta_m\end{bmatrix}=\left(X^TX\right)^{-1}X^TY \tag{2.5}$$

2.3.2 回归模型的显著性检验

首先建立待检假设

$$H_0:\beta_1=\beta_2=\cdots=\beta_m=0 \tag{2.6}$$

若能通过检验拒绝H_0，则 Y 与 m 个变量$x_1,x_2,\cdots,x_m$之间存在线性相关关系。记

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i,S_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2,Q=S_{yy}-U \tag{2.7}$$

选取统计量$F=\dfrac{U/m}{Q/\left(n-m-1\right)}$在$H_0$成立的条件下，$F$属于$F\left(m-1,n-m-1\right)$。

得到拒绝域：

$$W=\left\{F=\frac{U/m}{Q/(n-m-1)}\geq F_{\alpha}(m-1,n-m-1)\right\} \tag{2.8}$$

若$F>F_{\alpha}$，拒绝H_0，即 Y 与 m 个变量$x_1,x_2,\cdots,x_m$之间存在线性相关关系；否则，接受H_0，即 Y 与 m 个变量$x_1,x_2,\cdots,x_m$之间不存在线性相关关系。

2.4 模型的优化

2.4.1 回归模型中影响因子的优化

如果影响因子过多，包含不重要的因子。我们采用的办法是偏 F 检验法，以剔除不重要的因子，保留重要的因子。

检验假设$H_k:\beta_k=0,k=1,2,\cdots,m$，通常选取统计量$F_{\alpha}=\frac{\beta_k{}^2/\alpha_{kk}}{Q/(n-m-1)}$，其中$\alpha_{kk}$是矩阵$\left(X^TX\right)^{-1}$的主对角线上的第 k+1 个元素。

在H_k成立的条件下，F_k属于$F(m-1,n-m-1)$。得到拒绝域：

$$W=\left\{F_k=\frac{\beta_k{}^2/\alpha_{kk}}{Q/(n-m-1)}\geq F_{\alpha}(m-1,n-m-1)\right\} \tag{2.9}$$

若$F_k>F_{\alpha}$，拒绝H_k，即x_k与Y影响显著；否则，接受H_k，即x_k与Y影响不显著，可以剔除x_k。

2.4.2 预测问题

现在给出一个当前因子信息$(x_1,x_2,\cdots,x_m)=(x_{01},x_{02},\cdots,x_{0m})$，则预测量：

$$y_0=\beta_0+\beta_1x_{01}+\beta_2x_{02}+\cdots+\beta_mx_{0m} \tag{2.10}$$

预测量y_0的优劣取决于$|y_0-Y_0|$的大小。记

$$\overline{x_j}=\frac{1}{n}\sum_{i=1}^{n}x_{ij},\quad l_{ij}=\sum_{i=1}^{n}(x_{ki}-\overline{x_i})\left(x_{kj}-\overline{x_j}\right),i,j=1,2,\cdots,m, \tag{2.11}$$

$$L=\begin{bmatrix}l_{11}&\cdots&l_{1m}\\ \vdots& &\vdots\\ l_{m1}&\cdots&l_{mm}\end{bmatrix},L^{-1}=\begin{bmatrix}l_{11}{}'&\cdots&l_{1m}{}'\\ \vdots& &\vdots\\ l_{m1}{}'&\cdots&l_{mm}{}'\end{bmatrix}, \tag{2.12}$$

$$d^2=1+\frac{1}{n}+\sum_{i=1}^{n}\sum_{j=1}^{n}l_{ij}{}'\left(x_{0i}-\overline{x_i}\right)\left(x_{0j}-\overline{x_j}\right),\quad \hat{\sigma}^2=\frac{Q}{n-m-1} \tag{2.13}$$

可以证明当Y_0与$Y_1,Y_2,\cdots,Y_n$相互独立时，$\frac{y_0-Y_0}{d\hat{\sigma}}$满足$t(n-m-1)$。这样在显著性水平$\alpha$下可得到$Y_0$的预测区间：

$$\left[y_0-t_0(n-m-1)d\hat{\sigma},y_0+t_\alpha(n-m-1)d\hat{\sigma}\right] \tag{2.14}$$

2.5 模型中的理论问题

2.5.1 针对小抽样率下的信息收集

考虑到我国各地地区差异性比较大，各地的饮食习惯以及居民的整体消费水平的差异性，所以，需要对全国或者某个地区的某类食品进行摄入量的分析和统计，通过抽样调查,不仅仅需要获得对总体目标量特征的考察,同时也关注于总体内部的各子总体的状况。基于我国国情的这一特点，我们建立多层次估计的数学模型。

解决多层次目标量的估计问题可以有两种方法:一种是样本量的改变,如：层层抽样法、ABC 三级一套样本兼容法、样本追加策略等。这种思路存在着一些难以克服的缺憾，首先,没有样本单位入选的那些“小区域”,无法解决数据的推断问题；其次,在有样本单位入选的那些“小区域”,为了达到小区域内目标量估计的精度,必须有一定的样本量作保证,而这点恰恰与抽样调查的本质发生冲突。解决多层次需要的另一种方法是从估计方面入手,通过改进估计的技术,实现满足我们需要的目的。这里我们可以采用“小区域估计”法。小区域估计方法实际上就是使用间接估计的方法来解决上述多层次目标量的估计问题。它的中心思想就是在根据全国样本对省、市、县、乡的推断中,不再额外增加实际样本,而是广泛收集各种相关的辅助信息,例如以往的调查资料,普查资料、行政纪录等,通过挖掘这些变量间的关系,建立合理的模型,来达到增加“虚拟样本”的目的,从而对那些目标量估计精度不足的区域的估计值进行纠正,对那些没有被抽中的区域进行模拟和预测。

结合上述两种方法，我们分地区抽取样本，可以采用整群抽样法，即首先将全国各省市自治区直辖市按照不同的特点分成若干互不交叉、互不重复的地区，然后以地区为抽样单位抽取样本。通过对各个地区进行符合基本假设的抽样调查，获得有效的样本数据，然后采用小区域估计法获得该地区的整体情况。

(1)获取辅助信息

通过收集各个地区的普查资料，行政纪录和调查资料获取各个地区的如下信息：性别比例、各个年龄段（老、中、青少）人口比例、经济收入各阶段近似比例、体力与脑力劳动者比例、该地区各省市自治区直辖市的人口分布。

(2)根据辅助信息抽样获得有效的样本数据，其中有效性是指必须满足表 2 食物摄入量调查数据表的内容。

(3)对各个地区的抽样数据进行统计分析

对于某种食品，按地区作如表 3 的数据统计。

表 3 某地区某食品的日摄入量(单位:g)

		人均日摄入量	最高日摄入量	最低日摄入量	某右分位点	方差	人数
性别	男						
	女						
年龄	老年						
	中年						
	青少年						
季节	春秋						
	冬夏						
经济收入水平	0-500						
	500-2000						
	2000-5000						
	>5000						
劳动强度	体力						
	脑力						
	无业						

(4)根据辅助信息修正整体模型

分别比较抽样数据中的性别比例、各个年龄段（老、中、青少）人口比例、经济收入各阶段近似比例、体力与脑力劳动者与无业者比例与辅助信息中该地区的各项比例是否一致。

（a） 若一致或者基本一致，则说明抽取的样本能够很好的代表该地区的人群特征，统计的结果可以作为该地区的整体摄入量分布

（b） 若存在较大的差别，则进行样本以及估计方法的修正

2.5.2 样本修正

通过有限次复制抽样调查得到的样本，扩充样本数量，使得扩充后的样本总体的各项比例与辅助信息中的各项比例尽可能的接近。

用数学语言表述如下，设某地区抽取的样本总数为 N，其中男士、女士分别为 m 和 f，老年、中年、青少年分别为 a_1,a_2,a_3，c_1,c_2,c_3,c_4 分别表示四个不同经济收入水平的人数，l_1,l_2,l_3 分别表示体力、脑力劳动者与无业者的人数，显然：

$$m+f=a_1+a_2+a_3=c_1+c_2+c_3+c_4=l_1+l_2+l_3=N \tag{2.15}$$

通过多次数据复制，

$$\begin{cases} u_1 m+u_2 f=N' \\ \sum_{i=1}^{3} v_i a_i=N' \\ \sum_{i=1}^{4} q_i c_i=N' \\ \sum_{i=1}^{3} h_i l_i=N' \end{cases} \tag{2.16}$$

使得

$$\begin{cases} u_1 m : u_2 f = A + \delta_1 \\ v_1 a_1 : v_2 a_2 : v_3 a_3 = B + \delta_2 \\ q_1 c_1 : q_2 c_2 : q_3 c_3 : q_4 c_4 = C + \delta_3 \\ h_1 l_1 : h_2 l_2 : h_3 l_3 = D + \delta_4 \end{cases} \tag{2.17}$$

其中 N' 为扩充后的样本总数，$\delta_1, \delta_2, \delta_3, \delta_4$ 为无穷小量。通过调节 $\delta_1, \delta_2, \delta_3, \delta_4$ 的大小，使得上述方程组有解，并且 $\delta_1, \delta_2, \delta_3, \delta_4$ 尽可能的接近于 0。

第三部分 子模型 2—污染物分布模型

污染物分布模型是根据农药、化工等污染行业的污染物排放数据和食品卫生安全检测部门日常对水、农贸市场和大宗食品中污染物抽查数据以及进出口口岸的检测数据来估计各类食物中各种污染物的含量。

3.1 污染物测定方案

检测对象的选择原则：根据不同地区、不同季节市面上各类食品的流通量的多少情况，优先抽样检测流通量大的食品类型。同时结合食品摄入量模型中食品的分类原则，对这些类食品种几种危害面广、后果严重的污染物进行成分的检测。

检测实施方案：参照食品摄入量模型将全国分为同样的有限个区，每省为一个具体实施单位，下设多个调查点，如超市调查点和农贸市场调查点。对调查点的分类食物进行抽样检测污染物含量。

污染物的分类：

(1) 化学污染物：铅、镉、总汞、甲基汞、总砷、无机砷、二英及其类似物、氯丙醇、丙烯酰胺等；

(2) 农药残留：有机磷、有机氯、拟除虫菊酯、氨基甲酸酯、有机锡等杀虫剂以及某些杀菌剂、除草剂；

(3) 微生物污染(霉菌毒素)：黄曲霉毒素 B1，单增李斯特菌，沙门氏菌，空肠弯曲菌，副溶血性弧菌，出血性大肠杆菌，疯牛病，高致病性禽流感等。

3.2 污染物测定数据使用方案

食物的污染物检测方式有例行监测数据和偶然抽查数据，检测方法有符合性检验和监测性检验数据。为了减少工作量和监测时间的要求，我们采用根据污染物含量大于某一数值的部分样本数据和通过大约占数据总量 2%偶然抽查数据所获得的小于等于同一数值的部分样本数据来估计出这个污染物含量的整体分布。污染物检测数据构成如图 3 所示。

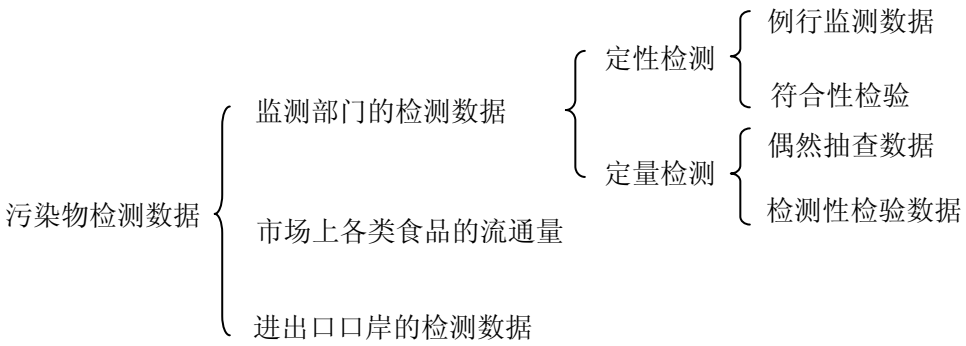

图 3 污染物监测数据构成

污染物检测数据表分为符合性检验数据表和监测性检验数据表，如表 4 和表 5 分别是各类食品中铅的监测性检验数据表和食源性致病菌检出数据表[3]。

表 4 各类食品中铅的检验数据表

食品品种	样品（份）	含量范围（μg/kg）	中位数（μg/kg）	国家标准（μg/kg）	超标（份）
大米	4	1.5～158.4	37.4	200	0
淡水鱼	17	1.5～9344.6	53.2	500	1
禽畜肉	17	1.5～616.4	77.6	200	3
猪肾	23	1.5～2456.6	123.7	500	6
皮蛋	17	2.0～5611.6	308.2	2000	3
茶叶	38	136.1～1735.1	846.9	5000	0

表 5 各类食品中食源性致病菌检出数据表

食品品种	样品（份）	沙门氏菌		单增李斯特菌	
		阳性(株)	(%)	阳性(株)	(%)
生猪肉	20	0	0	6	30
生牛肉	25	0	0	14	56
生鸡肉	36	1	2.8	11	30.6
散装熟肉	30	0	0	5	16.7
水产品	25	0	0	1	4
蔬菜	20	0	0	0	0
合计	156	1	0.6	37	23.7

3.2.1 可用的数据

通过抽样检测得到的定性数据（包括达标食品与超标食品的比例，各类食品之间合格率的大小关系）和不符合标准的那部分食品中污染物的含量，总抽样数据中约占 2%的那部分偶然抽查的详细样本数据。

使用方案：利用抽样中的 2%的偶然抽查数据来建立抽样出来的样本总体的分布情况。然后再利用样本中的超标的数据以及其他反应中国国情的辅助信息来修正这个模型，进而得到近似的各个地区的各种污染物分布模型。

3.2.2 对于定量数据(占数据总量 2%的偶然抽样数据)的使用

下面讨论如何通过大约占数据总量 2%的偶然抽样数据估算某污染物随机变量的整体分布。将占总数据 2%的偶然抽样数据，以及不合格（检测出污染物超标）数据作为定量数据 $Data_{2\%}$。

(1) 小样本抽样方法--Bootstrap 方法

Bootstrap 方法实质是一个再抽样的过程，即用现有的数据去模仿未知的分布，运用这种方法可以对参数进行区间估计或统计假设检验。设 $X=(x_1,\cdots,x_n)$ 为 $F(x)$ 样本，$\theta=\theta(F)$ 为总体分布的未知参数，F_n 为抽样分布函数，$\hat{\theta}=\hat{\theta}(F_n)$ 为 θ 的估计，记 $X^*=(x_1^*,\cdots,x_n^*)$ 为从 F_n 中抽样获得的再生样本，称其为 Bootstrap 子样。F_n^* 是由 X^* 所获得的抽样分布。记 $R_n=\hat{\theta}(F_n^*)-\hat{\theta}(F_n)$，称 R_n 为 T_n 的自助统计量，利用 R_n 分布（在给定 F_n 之下）去近似 T_n 的分布。

Bootstrap 方法是一种非参数统计方法，其目的是用现有的资料去模仿未知的分布，通过再抽样将小样本问题转化为大样本问题，适用于小样本条件下的统计推断。

(2) 小样本评估方法

正态分布总体的小样本的最优概率密度拟合是 t 分布。

引理 1：若 $Y_1\sim N(0,1)$，$Y_2\sim\chi_n^2$，且 Y_1 与 Y_2 相互独立，则

$$\frac{Y_1}{\sqrt{Y_2/n}}\sim t_n \tag{3.1}$$

。

引理 2：若 $X\sim N(\mu,\sigma)$，则

$$\frac{(n-1)S^2}{\sigma^2}\sim\chi_{(n-1)}^2, \tag{3.2}$$

定理 1：当 $X\sim N(\mu,\sigma)$ 时，

$$\sqrt{\frac{n}{n+1}}\frac{X-\overline{X}}{S}\sim t_{n-1}, \tag{3.3}$$

其中 $\overline{X}$ 为样本均值。

证明：由 $E(X-\overline{X})=0$，$D(X-\overline{X})=\sigma^2+\frac{\sigma^2}{n}=\frac{1+n}{n}\sigma^2$ 可知，

$$(X-\overline{X})\Big/\sqrt{\frac{1+n}{n}\sigma^2}\sim N(0,1) \tag{3.4}$$

另有，$\overline{X}$ 与 S^2 相互独立，且 $\frac{X-\overline{X}}{S}=\sqrt{\frac{1+n}{n}}\frac{(X-\overline{X})\Big/\sqrt{\frac{1+n}{n}\sigma^2}}{\sqrt{(n-1)S^2/(\sigma^2\bullet(n-1))}}$。

由引理 1 和引理 2，有

$$\sqrt{\frac{n}{n+1}}\frac{X-\overline{X}}{S}\sim t_{n-1}, \tag{3.5}$$

其中 n 为样本量，$n-1$ 为 t 分布的自由度。（证毕）

由定理 1 我们可以知道 $\frac{X-\overline{X}}{S}\not\sim N(0,1)$，但是由 $\lim_{n\to 0}\sqrt{\frac{n}{n+1}}=1$，有

$$n\to\infty, t_n\to N(0,1)。 \tag{3.6}$$

即随着样本量增大 t 分布逼近标准正态分布，再由 t 分布的分布函数连续性可知，

$$\lim_{n\to 0}t_{n-1}\left(\sqrt{\frac{n}{n+1}}x\right)=\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-x^2/2}dx。 \tag{3.7}$$

所以，当样本量足够大时，可以用样本特征量代替母体参数估计可靠度。但是实际情况下样本是很少的，这时我们由式 2.20 可知

$$R=P(X\le x)=t_{n-1}\left(\sqrt{\frac{n}{n+1}}\frac{x-\overline{X}}{S}\right) \tag{3.8}$$

t_{n-1} 为自由度为 $n-1$ 的 t 分布的分布函数，x 为性能参数的容许上限。

根据样本 $x_1,x_2,\cdots,x_n$ 可计算出 $\sqrt{\frac{n}{n+1}}\bullet\frac{x-\overline{X}}{S}$ 的值，然后根据自由度为 $n-1$ 的 t 分布表查表，即可得到可靠度。

3.3 污染物分布模型

下面具体讨论对于定量数据 $Data_{2\%}$ (占数据总量 2%的偶然抽样数据)的具体应用。

(1) 取 $Data_{2\%}$ 对数，统计频数，如果其频数满足正态分布，则采用对数正态分布法建立模型。

令样本数据为 u，且

$$\begin{cases} x=\lg(u) \\ f_x(x)=\dfrac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/2\sigma^2} \end{cases}, \tag{3.9}$$

显然，

$$F_x(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/2\sigma^2}dy\equiv G\left(\frac{x-\mu}{\sigma}\right) \tag{3.10}$$

这里，$G(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-y^2/2}dy$。

(2) 如果 $Data_{2\%}$ 的对数不满足正态分布，则采用 Nakagami-m 分布[4]近似表示。

如果随即变量 x 的概率密度函数是

$$f_x(x)=\begin{cases}\dfrac{x}{\sigma^2}e^{-x^2/2\sigma^2} & x\geq 0\\ 0 & x<0\end{cases} \tag{3.11}$$

通过引入参量 m，从瑞利分布可以导出类更一般的分布——Nakagami-m 分布，其概率密度函数是

$$f_x(x)=\begin{cases}\dfrac{2}{\Gamma(m)}\left(\dfrac{m}{\Omega}\right)^m x^{2m-1}e^{-x^2/2\Omega} & x>0\\ 0 & x\leq 0\end{cases} \text{。} \tag{3.12}$$

通过调整 m 可以控制密度函数的拖尾， m 值越大，概率密度函数拖尾衰减越快。

第四部分 综合模型—风险评估模型

风险就是度量一个不利事件发生的可能性及其不利结果。风险评估模型是利用人群食物摄入量模型和污染物分布模型的数据对全国、某个地区某类食品的安全状况做出评价，对可能出现的食品安全事件给出预警。

4.1 风险评估框架

风险评估框架如图 4 所示。

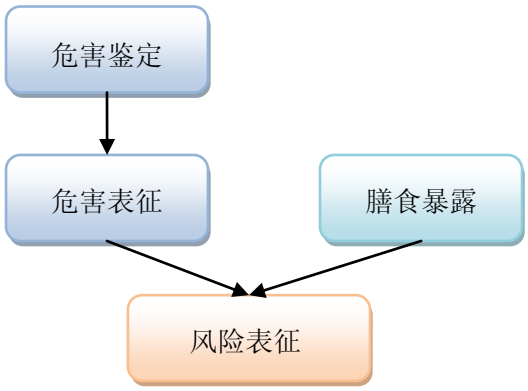

图 4 风险评估框架

危害鉴定与危害表征可由相关部门进行检测和鉴定，我们主要对膳食暴露加以研究与分析。我们首先分析某个地区某个季节的膳食暴露(即某段时间污染物的摄入量)情况。广义而言，代表这种膳食暴露的模型可以用下列公式来笼统的表示：

膳食暴露=食物摄入量×食物中某种污染物含量

4.2 地区膳食暴露风险评估模型

我们考虑单个地区某个季节某特定人群的食物安全情况。关键的问题是如何将人群摄入量的数据和食物中某种污染物含量与浓度的数据耦合起来。下面我们主要考虑以下两种方法风险评估模型。

4.2.1 利用食物摄入量模型和污染物模型中的抽样数据直接进行计算和评估分析

考虑某种污染物的情况，由抽样数据整理统计得到抽样人群中每个人平均每天对于该种污染物的摄入量。不妨设抽样的总人数为 N， X_i（$i=1,2,\cdots,N$）表示样本个人 i 对于某种污染物的日摄入量，则

$$X_i=\sum_{j=1}^{n}x_{ij}u_j \quad 。 \tag{4.1}$$

其中 x_{ij} 表示第 i 个人第 j 类食物的摄入量， u_j 表示第 j 种食物中含该种污染物的浓度，其中 n 表示的是摄入量统计和污染物检测过程中均调查的食物种类，可以是一种食品作为一类，也可以是多种同类型的食品归为一类食物，视调查的方案而定。

因为在污染物分布抽样检查中，对于未超标（ $X_i\le PTDI$ ）的食品不能给出污染物具体的浓度，因此，我们可以假定样本中污染物超标的个人摄入量数据是已知的，而且达标与超标的人数比例是可以统计出来的（其中 $PTDI$ 即是 Provisional Tolerable daily Intake，表示某种污染物的人均日安全摄入量，是一种常用的安全摄入水平指标）。

下面引入 Plug-in 估计和 TE 估计[5]两种方法来进行风险评估分析。

Plug-in（PI）估计：

$$P=\frac{N(X_i>PTDI)}{N} \tag{4.2}$$

其中 $N(X_i>PTDI)$ 表示 $X_i>PTDI$ 的样本数目。通过计算这个概率，我们可以预估该种污染物超标的概率 P。这种估计十分简单直观，但是不够精确。然而，对于各个地区先作 PI 估计能够，选出那些 P 的值比较大的地区，对我们的预警具有很大的指导作用，可以指导政府提高对这些地区的食品安全的警惕性，加大对这些地区的作进一步详细抽样调查的力度。对于 P 值比较高的地区我们可以进行如下比较精确的 TE 估计。

Tail Estimation(TE)估计：即主要考虑超标部分的数据的分布函数估计。因为超标的部分污染物含量较大，而概率较低，并且含量越大，出现的概率越低，所以在以污染物含量为变量的概率分布图中处于均值点右侧尾部。因此称为基于尾部数据的估计法（TE 估计）。不妨设 $X_1,X_2,\cdots,X_N$ 是独立同分布的，对其进行由小到大的排序，得到新的样本序列 $X_{1N},X_{2N},\cdots,X_{NN}$。其中 X_{NN} 表示这 N 个样本中的最大值。对于 X_{NN}，当 $N\to\infty$ 时，其分布 R_γ 一般可以用 Gumbel 分布（$\gamma\to 0$）、Fréchet 分布（$\gamma>0$）或者 Weibull 分布（$\gamma<0$）来逼近。 这是根据极限值理论中的主要理论（Fisher Tippett 定理）得到的。那么如何对 R_γ 进行调整和修正，使得尾部暴露的分布更接近于现实的情况呢？我们利用 Pareto 法则。对

于处于尾部的变量（$x > PTDI$），假设$1-F^0(x)=Cx^{-1/\gamma}$，其中$F^0(x)$为污染物含量小于等于x的概率。通过选取足够大的k，采用如下办法能够估计出参数C和γ：

$$\begin{cases}\gamma(k)=Hill_{k,N}=\dfrac{1}{k}\sum_{i=1}^{k}\log\dfrac{X_{N-i+1,N}}{X_{N-k,N}}\\ C(k)=\dfrac{k}{N}(X_{N-k,N})^{1/Hill_{k,N}}\end{cases} \tag{4.3}$$

其中$Hill_{k,N}$也成为 Hill 估计因子。除了引入一个足够大的k来估计参数C和γ之外，还可以引进一个调节函数对R_γ进行修正。反映在分布函数上也就是：

$$1-F^0(x)=Cx^{-1/\gamma}L(x) \tag{4.4}$$

其中调节函数$L(x)$可以选择任意满足如下性质的函数：

$$\lim_{x\to\infty}\frac{L(tx)}{L(x)}=1,\qquad \forall t>0. \tag{4.5}$$

如取$L(x)=1+Dx^{-\beta}$，其中 D 是非负的，$\beta>0$。

通过上述的调整与修正，我们能求得$F^0(x)$，也就得到相应的污染物摄入量的各种评估指标。如相应的 99.999%右分位点$u_{0.99999}=(F^0)^{-1}(0.99999)$。

4.2.2 利用食物摄入量模型和污染物模型中抽象出来的函数关系与概率分布进行计算和评估分析

假定抽样食物种类为 n，污染物种类为 m。记某个地区某个季节的食物摄入量函数为

$$G(x)=(g_1(x),g_2(x),\cdots,g_n(x))^T \tag{4.6}$$

其中$x=(x_1,x_2,x_3,x_4,x_5)^T$为五维向量，其中各个分量分别代表人的性别、年龄、经济水平、劳动强度、体重等变量因子。

各种污染物在第 i 类食物中的分布函数为

$$F_i(u)=(f_{i1}(u),f_{i2}(u),\cdots,f_{im}(u))^T \tag{4.7}$$

(1) 点估计 对于每个$g_i(x),f_{ij}(u)$分别用固定的值去逼近。即，选取适当的$\tilde{g}_i,\tilde{f}_{ij}$去替换函数$g_i(x),f_{ij}(u)$，从而得到这 m 种污染物日摄入量水平：

$$\begin{aligned}&H=(h_1,h_2,\cdots,h_m)^T\\ &h_j=\sum_{i=1}^{n}\tilde{g}_i\tilde{f}_{ij},\quad j=1,2,\cdots,m\end{aligned}\ 。 \tag{4.8}$$

关于 $\tilde{g}_i, \tilde{f}_{ij}$ 的取值，可以是相应的函数的平均值，最大值，或者是处在某个百分位点的值，因此能得到这 m 种污染物日摄入量的各种评价水平——均值水平 H_{ave}、最值水平 $H_{\max}$、α 百分位点水平 $H_{\alpha\%}$。这种估计比较简单，操作比较容易，具有一定的指导性。但是误差比较大。

(2) 单一分布[6]

(a) 对于每个 $f_{ij}(u)$ 仍然用固定的值 $\tilde{f}_{ij}$ 去逼近，而考虑食物摄入量的函数关系。此时我们能得到污染物日摄入量关于 x 的函数：

$$H(x) = (h_1(x), h_2(x), \cdots, h_m(x))^T$$
$$h_j(x) = \sum_{i=1}^{n} g_i(x)\tilde{f}_{ij}, \quad j = 1, 2, \cdots, m \tag{4.9}$$

结合人群的各种统计特征，如某个地区的男女性别比例、人口的年龄结构、各种劳动强度比例、经济收入水平统计等等辅助信息可以得到关于 x 的一个近似的分布规律，然后对 $H(x)$ 进行概率统计，便能得到相应的每种污染物日摄入量的各种评价水平分析。

(b) 对于每个 $g_i(x)$ 仍然用固定的值 $\tilde{g}_i$ 去逼近，而考虑污染物的随机分布。例如可用 Pareto 指数估计。

在 Pareto 指数估计将食物风险定义如下

$$\text{风险}=\text{食物摄入量}\times\left(\text{污染物分布}\right)^{\alpha}。$$

若记食物摄入量模型为 S，污染物分布模型为 W，则有

$$F = S \times (W)^{\alpha}。 \tag{4.10}$$

这里面的 α 称为 Pareto 指数。当食物摄入量数据和污染物分布抽样数据的对象不一致时，不妨假设这两个抽样调查是完全独立进行的，那么，我们将这个等式两端同时取对数，之后就有：

$$\begin{array}{ccccc} \ln F & = & \ln S & + & \alpha \ln W \\ \downarrow & & \downarrow & & \downarrow \\ y & = & x_1 & + & \alpha x_2 \end{array}。 \tag{4.11}$$

在某一个地区的某一个季节下，x_1 就可以当作一个定值了，我们现在来求 α。我们在测定污染物分布时的符合性检验数据的合格与不合格的比值，有一个比例 $n_1 : n_2$（这里的数据可是按照污染分布模型中那个扩充方法扩充后的数据，即是整体数据）。那么，在合格与不合格的这个临界点处的不合格风险概率应该近似等于 $n_2/(n_1+n_2)$，就是在这一点处 y 的值。x_1 已经是确定的值了，现在就剩下 x_2 了。对于 x_2， 不合格时，污染物的含量全部都

是定值，分成 n 个区间，取第 i 区间中的频数 p_i（$\sum_{i=1}^{n} p_i = 1$），以及第 i 个区间的均值 a_i。这时 x_2 的取法有三种：

Ⅰ.合格与不合格临界点的值（这个值太小，预警敏感度太高）；

Ⅱ.所有不合格数据中超标最高的那个值（这个值太大，预警灵敏度过低）；

Ⅲ.取数学期望 $E = \sum_{i=1}^{n} a_i p_i$（和上面两个比起来，这个合理）。

取完 x_2，代进去式 4.11 就可以求出 α，从而得到整个模型中的分布函数，于是就能求出我们需要的百分位点的值。

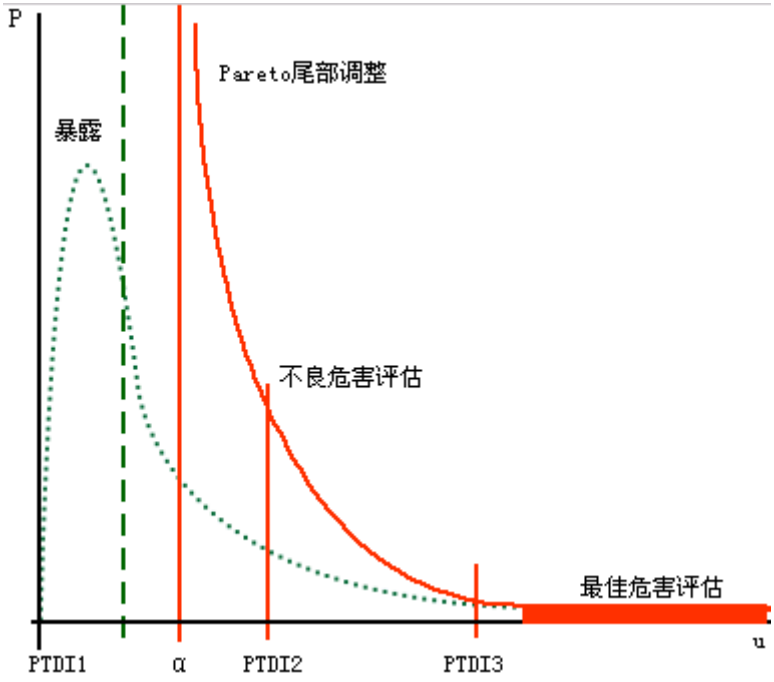

图 5

图 5 中点绿曲线就是把 x_2 用第一种方式取的，这个实红曲线就是把 x_2 用第二种方式取的，我们最希望 x_2 的取值越接近真实越好，所以用数学期望比较好。当食物摄入量数据样本不能很好的拟合成函数或者污染物分布数据样本抽象出来的概率分布与现实情况存在太大差异时，我们可以考虑用这种单一分布的处理方法。

(3) 双随机分布

对于每个 $g_i(x), f_{ij}(u)$ 均不考虑用固定的值去逼近，而是看成函数来处理。此时污染物日摄入量是关于 x 和 u 的函数，即

$$H(x,u) = (h_1(x,u), h_2(x,u), \cdots, h_m(x,u))^T$$

$$h_j(x,u) = \sum_{i=1}^{n} g_i(x) f_{ij}(u), \quad j = 1,2,\cdots,m \qquad (4.12)$$

。

利用食物摄入量模型和污染物分布模型建立的函数关系和概率分布，我们能得到相应的 $H(x,u)$ 的具体表达式，从而直接对 $H(x,u)$ 进行概率统计和风险分析，得到相应的每种污染物日摄入量的各种评价水平分析。这个估计对前两个模型中的 $g_i(x), f_{ij}(u)$ 的准确率要求比较高，而准确率又需要样本数来保证，因此操作比较复杂，但是预警能力比较强。因为时间关系，我们这里对这种估计只是提出一个思路，不准备用实例来进行验证。

4.3 全国膳食暴露风险评估模型

当各个省市的分布于全国不同，怎么得到全国的总体分布？ 我们采用聚类分析法解决这一问题。聚类分析的内容很丰富，系统聚类法开始每个对象自成一类，然后每次将最相似的两类合并，合并后重新计算新类与其他类的距离或相近性测度。这一过程一直继续直到所有对象归为一类为止。并类的过程可用一张谱系聚类图描述。最后再将比例回代就可以求出每一个地区在全国的权重。将这权重乘到对应该地区的数据上，就是最后归依与一类的信息数据。再利用这些加权的信息来建立全国的模型。

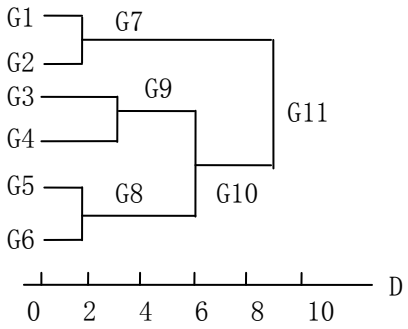

图 6 聚类图

例如我们得出的第一个摄入量模型，我们知道对于每一个地区的摄入量 y_i 是由各个因素组即

$$y_i = \beta_{i0} + \beta_{i1}x_{i1} + \beta_{i2}x_{i2} + \cdots + \beta_{im}x_{im} \text{。} \tag{4.13}$$

造成全国的分布与地区不同的主要原因也是这些影响因子因地域的不同而显著的不同。所以，我们这些地区之间的子总体之间的距离应该以这些影响因子 β_j 的差别来定义，可以采用如下任意一种距离来进行分析。

$$d(y_i - y_{i'}) = \sum_{j=0}^{m} \left|\beta_{ij} - \beta_{ij'}\right| \text{（绝对值距离）} \tag{4.14}$$

$$d(y_i - y_{i'}) = \sqrt{\sum_{j=0}^{m} \left|\beta_{ij} - \beta_{ij'}\right|^2} \text{（欧式距离）} \tag{4.15}$$

$$d(y_i - y_{i'}) = \frac{1}{m}\sum_{j=0}^{m} \frac{\left|\beta_{ij} - \beta_{ij'}\right|}{\left(\beta_{ij} - \beta_{ij'}\right)} \text{（兰氏距离）} \tag{4.16}$$

第五部分 模型检验

5.1 数值模型检验

我们根据文献[3]中污染物监测数据，对北京市东城区、某一季节下抽样 20 人，每人每日摄入量（千克）与铅含量对我们建立的食物风险评估模型进行检验。

5.1.1 子模型 1 数据

表 6　50 人的具体详细数据

50 人 样本数据　　单位：克							
样本	大米	淡水鱼	禽畜肉	猪肾	豆制品	皮蛋	茶叶
P1	158.8776	0.9147	149.4625	9.7569	1.7899	15.7703	1.0192
P2	221.1388	48.8923	23.5938	1.5871	0.8756	3.8683	1.0147
P3	126.8789	15.7835	140.4359	2.9417	5.2027	5.718	0.0451
P4	540.8232	1.4998	65.8791	4.145	1.4555	0.9139	0.455
P5	185.1914	22.1883	255.6613	8.3743	9.7549	1.2618	0.657
P6	121.0502	21.5806	53.4945	9.6115	5.4155	17.0042	2.1167
P7	206.3987	108.8213	320.2525	2.1833	0.2259	8.4949	0.7855
P8	217.8474	38.1854	43.9198	0.2249	19.169	4.415	2.5436
P9	294.5191	51.6871	114.9901	12.083	5.8363	7.2341	1.5188
P10	108.5916	85.0526	88.6765	1.5495	0.9624	10.2253	1.5298
P11	168.282	77.0639	28.3386	0.9379	4.1186	7.8615	0.4993
P12	159.465	1.9749	103.5336	4.7385	6.6169	17.6422	0.6911
P13	189.9817	36.1432	36.4878	2.6285	24.9583	9.7248	1.1255
P14	106.4761	25.0769	134.1987	5.578	3.6481	15.3139	0.538
P15	131.4569	28.9266	100.8872	7.9616	21.1502	20.6672	0.4426
P16	272.3077	61.7622	21.3874	5.8742	4.4376	1.2225	1.5397
P17	332.6586	11.6788	29.9264	2.4257	11.2486	9.624	0.3086
P18	179.3637	47.6685	25.5849	8.2274	8.1322	15.01	2.069
P19	330.5694	19.1187	19.0826	2.2712	24.9002	14.8526	0.8796
P20	189.5914	16.1225	7.906	5.0438	13.133	18.338	0.6168
P21	301.369	16.7572	69.9851	10.247	12.8334	11.3231	1.8798
P22	245.3769	16.2646	79.654	3.0101	5.2641	7.574	0.7093
P23	107.4159	4.3015	80.1284	0.0893	12.7369	0.9991	1.5647
P24	236.1175	29.9375	0.7599	8.0521	2.1595	25.6805	0.8315
P25	116.0432	24.2595	72.681	6.1938	2.8317	15.1616	0.1733
P26	461.5931	6.5567	149.0902	3.4179	20.8419	11.0077	0.4162
P27	295.1052	16.9686	87.0335	3.9036	8.7029	10.1732	0.1965
P28	276.6143	80.1655	26.5675	2.6549	11.3335	29.6172	0.6631
P29	223.928	1.8642	156.6695	10.6725	27.0031	8.0299	1.0927
P30	341.1063	8.9598	39.4636	1.7718	5.5448	3.1191	1.3078

P31	263.6638	18.4598	14.3855	1.1585	12.0596	4.4679	2.6268
P32	174.3144	12.4621	233.4247	6.4399	34.0471	15.8226	2.4921
P33	191.1094	25.5388	135.7539	0.0674	11.7672	7.837	2.5312
P34	391.5546	112.3791	181.5689	6.6667	3.9263	10.7462	0.2971
P35	114.5621	2.0503	110.8422	2.7817	11.4858	8.8305	0.5531
P36	294.5902	5.1226	14.2115	3.778	17.1896	2.1519	0.1311
P37	302.4132	12.4289	111.2757	4.5593	0.8899	13.0161	0.0021
P38	412.4217	59.4173	55.9264	6.8584	9.0337	1.643	0.4759
P39	448.7425	13.4194	47.837	1.2279	24.7984	0.1873	0.2143
P40	394.3123	34.2742	67.9385	0.5942	0.8346	11.2619	1.0839
P41	326.0692	26.3317	28.4967	1.9235	10.6425	17.737	0.5155
P42	138.4368	3.7737	133.2935	0.3508	23.6726	4.3432	1.5014
P43	135.8078	9.7379	72.398	2.8915	15.4514	5.7335	0.3011
P44	155.3499	43.5454	66.36	2.3465	20.475	3.8609	2.9114
P45	203.0186	7.9289	19.8443	6.4984	3.307	22.5309	0.0591
P46	275.9767	5.2627	179.4868	8.174	15.3677	8.1663	0.5198
P47	180.4257	39.5076	138.7673	3.514	33.8133	2.8008	2.288
P48	294.7765	29.6817	19.7768	4.0363	5.6296	14.8739	0.8827
P49	527.0766	84.1014	146.9328	5.1373	12.7978	1.5079	0.5025
P50	141.2719	31.7869	36.6493	6.4725	1.9034	2.1566	0.2132

多元线性回归模型为：

$$\begin{aligned} Y = {} & 13.0903 + 7.2683x_1 + 1.7335x_2 + 4.3723x_3 + 0.2085x_4 \\ & + 0.5836x_5 + 0.4602x_6 + 0.0511x_7 \end{aligned} \tag{5.1}$$

其中 $x_1, x_2, \cdots, x_7$ 代表各个因素权重，不妨设 $x_1, x_2, \cdots, x_7 \in [1, 2, \cdots, 50]$。

最后通过线性回归求得该地区在该季节的人均摄入量，单位为千克，如表 7。

表 7 线性回归求得人均摄入量

大米	0．2231	y1
淡水鱼	0．0342	y2
禽畜肉	0．0826	y3
猪肾	0．0042	y4
豆制品	0．0110	y5
皮蛋	0．0094	y6
茶叶	0．0010	y7

5.1.2 子模型 2 数据

表 8 子模型 2 数据

食品品种	国家标准 $\mu g/kg$	符合性检测		监测性检验		
		样本量	合格份数	样本量	含量范围 $\mu g/kg$	超标份数

大米	200	20	20	4	1.5-158.4	0
淡水鱼	500	850	40	17	1.5-9344.6	1
禽畜肉	200	850	113	17	1.5-616.4	3
猪肾	500	460	115	23	1.5-2456.6	6
豆制品	1000	320	12	16	2.0-299.5	0
皮蛋	2000	850	120	17	2.0-5611.6	3
茶叶	5000	150	4	30	136.1-1735.1	0

食品品种	污染物（铅）分布 $\mu g/kg$
大米	37.4
淡水鱼	53.2
禽畜肉	77.6
猪肾	123.7
豆制品	43.8
皮蛋	308.2
茶叶	846.9

5.1.3 点-点风险模型结果

$$X_i = \sum_{j=1}^{7} x_{ij} u_j = 21.3215 \mu g \tag{5.2}$$

表 9 点-点风险模型检验数据

样本	铅摄入量	样本	铅摄入量	样本	铅摄入量	样本	铅摄入量
P1	24.6009	P14	21.7569	P27	22.8599	P40	26.3449
P2	14.9918	P15	22.9413	P28	27.1882	P41	22.4159
P3	18.8752	P16	17.7361	P29	26.5381	P42	19.4167
P4	26.6637	P17	19.4062	P30	18.8313	P43	14.2728
P5	30.3563	P18	18.988	P31	16.2404	P44	18.1276
P6	18.2924	P19	21.5579	P32	34.5785	P45	17.4976
P7	41.9259	P20	15.9371	P33	24.131	P46	29.1725
P8	17.9771	P21	24.5106	P34	39.2736	P47	24.3416
P9	27.9587	P22	19.7636	P35	17.0339	P48	20.2186
P10	20.1529	P23	12.6708	P36	14.3879	P49	37.6767
P11	15.7363	P24	20.1945	P37	25.2228	P50	11.5485
P12	21.0039	P25	16.981	P38	25.0804		
P13	17.2317	P26	34.2638	P39	22.687		

世界发达国家儿童血铅＜60 微克/升为相对安全，国际血铅诊断标准≥100 微克/升为铅中毒，则通过该估计表明北京市东城区居民铅中毒的概率远小于 0.001%。在本风险评估中 99.999%的右分位点为 41.9259，即 P7。该分位点小于国际铅中毒含量的标准，所以通过该风险评估模型分析得到是该地区的铅中毒概率符合要求。

第六部分 模型的分析与评价

通过人群食物摄入量模型、污染物分布模型、风险评估模型三部分的具体分析我们建立了一套比较完整的膳食暴露评估数学模型。在第一个模型中，我们首先制定了一个较好的抽样调查方案，分区设点进行抽样，食品分类进行统计。同时考虑到人群的性别、年龄、经济收入、劳动强度等因素。然后我们利用辅助信息增设虚拟样本来扩充样本数，弥补因为人力、物力、时间等因素带来的低抽样率的不足。同时，利用多元线性回归技术得到了食物摄入量函数。在第二个模型中，我们利用抽样中的 2%的偶然抽查数据通过来建立抽样出来的样本总体的分布情况。然后再利用样本中的超标的数据以及其他反应中国国情的辅助信息来修正这个模型，进而得到的各个地区的各种污染物近似的分布模型。对于 2%的小样本数据，我们进行多次再抽样这种 Bootstrap 方法获得大样本。鉴于污染物含量的分布显然不是正态分布而且很可能是左偏态的这一特征，我们引入了信道处理中的 Nakagami-m 分布来逼近。在第三个模型中，我们分两种思路进行探讨：一种思路是直接利用上述两个模型中的抽样数据进行计算获得抽样样本中每个人的各种污染物日摄入量，然后对于超标部分的数据进行 PI 估计，提供各个地区的超标概率，筛选出高风险地区进行更精确的风险评估——TE 估计。TE 估计主要是利用极限值理论对超标数据进行尾部分布的逼近，然后用 Pareto 法则对这一分布进行进一步的修正，得到与现实情况相似的概率分布函数，进而求出我们比较关注的 99.999%右分位点的值。另一种思路是利用前两个模型中抽象出来的函数和概率分布，进行点评估、单一分布、双随机分布三种估计方法来得到相应的风险评估指标。点估计比较简单，完全不用考虑数据的不匹配性，易于操作，能对我们的风险评估提供一定的参考信息，但是精确度不高；单一分布的精度较高，而且也能克服数据的不匹配的缺憾，但是对于数据的随机性与不确定的应变能力不够强。双随机分布是一类比较理想的估计法，它考虑了各参数的变化性与不确定性，当前两个模型的精度比较高时，这种估计能够很好地提供风险评估的各种指标。最后我们选择了一个简单的实例来说明第一种思路中的 99.999%右分位点的求法。

查考文献

[1] 2007 中国卫生统计提要数据, http://www.moh.gov.cn/open/2007tjts/P76.htm , 2007 年 10 月 20 日.

[2] 李艳平,何宇纳,翟凤英等, 称重法、回顾法和食物频率法评估人群食物摄人量的比较, 中国预防医学杂志,第 40 卷第 4 期, P273-279,2006 年 7 月.

[3] 余晓辉,赵春玲,陈旭东等, 北京市东城区 2005 年食品污染物监测结果分析, 中国初级卫生保健,第 20 卷第 8 期,P62-63,2006 年 8 月.

[4]Nakagami-m. A.帕普里斯, S.U.佩莱著, 保铮等译, 概率、随机变量与随机过程,西安, 西安交通大学出版社,2004, P70.

[5] J.Tressou, A.Crepet,P.Bertail,etc, Probabilistic exposure assessment to food chemicals based on extreme valuetheory. Application to heavy metals from fish and sea products, Food and Chemical Toxiclogy,42(2004), P1349-1358.

[6]罗伟,陈冬东,唐英章等, 论食品安全暴露评估模拟模型,食品科技,No.2,2007,P22-23.